Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models

Marijana Zekic-Susac
University of J.J. Strossmayer in Osijek, Faculty of Economics in Osijek
Gajev trg 7, Osijek, Croatia
marijana@efos.hr

Natasa Sarlija
University of J.J. Strossmayer in Osijek, Faculty of Economics in Osijek
Gajev trg 7, Osijek, Croatia
natasa@efos.hr

Mirta Bensic
University of J.J. Strossmayer in Osijek, Department of Mathematics
Gajev trg 6, Osijek, Croatia
mirta@mathos.hr

Abstract: *The paper compares the models for small business credit scoring developed by logistic regression, neural networks, and CART decision trees on a Croatian bank dataset. The models obtained by all three methodologies were estimated; then validated on the same hold-out sample, and their performance is compared. There is an evident significant difference among the best neural network model, decision tree model, and logistic regression model. The most successful neural network model was obtained by the probabilistic algorithm. The best model extracted the most important features for small business credit scoring from the observed data.*

Keywords. credit scoring modeling, decision trees, logistic regression, neural networks, small business

1. Introduction and previous research

Previous research on credit scoring was mostly focused on large company or consumer loans, while small business credit scoring, especially on croatian datasets, is poorly or not known. Feldman's research [6] showed that variables that effect small business loans differ from those that effect company loans. First research of NNs in credit scoring was done by Altman et al. [1] who compared linear discriminant analysis (LDA) and LR with NNs, in predicting company distress. Their results showed that LDA was the best in perfomance by recognizing 92.8% healthy and 96.5% unsound firms, while NN recognized 91.8% healthy and 95.3% unsound firms. LDA was also found as superior than NNs, genetic algorithms and decision trees in credit assessment by Yobas et al. [18], while the results of Galindo and Tamayo [9] showed that CART decision-tree models outperformed NNs, the k-nearest neighbor and probit algorithms on a mortgage loan data set. West [18] found some NN algorithms (mixture-of-experts, radial basis function NN, multi layer perceptron), and LR as superior methods, while others were inferior (learning vector quantization and fuzzy adaptive reasonance, as well as CART decision trees).

Since most authors except West [18] test a single NN algorithm, mostly the multi-layer perceptron such as backpropagation, we were challenged to compare the efficiency of more of them using additional optimizing techniques. Furthermore, general NN's best accuracy in classifying bad credit applicants emphasizes the potential efficiency of this methodology to recognize some hidden relationships among important features that traditional methods are not able to discover.

We assume that, in addition to methodology, specific economic conditions of transitional countries such as difficult access to the turnover capital, legal and law limitations, undeveloped infrastructure, high transactional costs, as well as high loan interest rates, also influence small business modeling [17]. Therefore, our objective was to find the best model that will extract important features for small business credit scoring in such specific environment. To do this, we investigated the predictive power of logistic regression (LR) in comparison to neural networks (NNs) and CART decision trees (CART) on the observed dataset.

2. Methodology

2.1. Logistic regression

Logistic regression modeling is widely used for analyzing multivariate data involving binary responses that we deal with in our experiments. As we had a small data set along with a large number of independent variables, to avoid overestimation we included only the main effects in the analysis. In order to extract important variables we used forward selection procedure available in SAS software, with standard overall fitting measures. Since the major cause of unreliable models is in overfitting the data [5], especially in datasets with relatively large number of variables as candidate predictors (mostly categorical) and relatively small data set such as the case in this experiment, we can not expect to improve our model due to addition of new parameters. That was the reason for investigating if some non-parametric methodologies, such as neural networks and decision trees can give better results on the same data set.

2.2. Neural Networks and CART

Since there are no defined paradigms for selecting NN architectures, we test four NN algorithms: backpropagation, radial-basis function (RBFN), probabilistic and learning vector quantization (LVQ) by NeuralWorks Professional II/Plus software.

First two algorithms were tested using both sigmoid and tangent hyperbolic functions in the hidden layer, and the SoftMax activation function in the output layer [13]. The learning is improved by Extended Delta-Bar-Delta rule, and a simulated annealing procedure [13]. The learning rate and the momentum were initially set to 0.5 and exponentially decreased during the learning process according to the preset transformation point [13]. Overtraining is avoided by a cross-validation procedure which validates the network result after each 1000 iterations and saves the best one [13]. The initial training sample (75% of total sample) was divided into two sub-samples: 85% for training, and 15% for cross-validation. After training and cross-validating the network on maximum 10000 iterations, all the NN algorithms were tested on the out-of-sample data (25% of the total sample).

The probabilistic neural network (PNN) algorithm was chosen as one of the stochastic-based networks, developed by Specht [12]. In order to determine the width of Parzen windows (σ parameter), we follow a cross-validation procedure for optimizing σ suggested by Masters [12].The last tested was the LVQ algorithm. Improved versions called LVQ1 and LVQ2 were used in our experiments [13].

The topology of networks consisted of an input layer, a hidden or a pattern layer (or Kohonen layer in LVQ), and an output layer. The number of neurons in input layers varied due to the suggested forward modeling strategy; the number of hidden neurons was optimized by the pruning procedure, while the output layer consisted of two neurons representing classes of bad and good credit applicants.

In order to find the best NN model, we introduce a slightly modified Refenes et al. [16] approach of introducing input variables into the model. Our approach is totally based on a nonlinear variable selection, starting from one input variable and gradually adding them in a forward procedure. Another feature that differs our strategy from [16] is in testing more than one NN algorithm. The procedure results in the best four NN models obtained by four algorithms. The best overall model is selected on the basis of the best total hit rate. The advantage of such nonlinear forward strategy allows NN to discover nonlinear relationships among variables that are not detectable by linear regression.

As the third method, we tested the CART decision tree, because of its suitability for classification problems, and as one of the most popular decision tree methods. The approach we use in our experiment was pioneered in 1984 by Breiman, et al. [19], and it builds a binary tree by splitting the records at each node according to a function of a single input field. The evaluation function used for splitting in CART is the Gini index [2] It considers all possible splits in order to find the best one. The winning sub-tree is selected on the basis of its overall error rate when applied to the task of classifying the records in the test set [4]. We performed the CART classification procedure using the Statistica software tool.

3. Data description and sampling procedure

Since other researchers have found both personal and business activities relevant in small business credit scoring systems [3],[6],[7],[8], we incorporate their findings in our variable selection. The lack of credit bureau in Croatia prevented us from collecting the information about personal repayment history of a small business owner. Data was collected randomly in a Croatian savings and loan association, and the sample size consisted of 166 applicants. The dataset resulted in the total of 31 variables, furtherly reduced to 20 using the information value proposed by Hand and Henley [10]. Descriptive statistics, separately for Good (G) and for Bad (B) applicants for used variables is presented in Table 1, where variables marked with "*" were found significant in the best model.

Table 1. Input variables with descriptive statistics

Variable	Descriptive statistics
Main activity of the firm *	Textile production and sale (G: 11.32% B:16.98%); Cars sale (G: 7.55% B: 9.43%); Food production (G: 25.43% B: 13.21%); Medical, intelectual services (G: 18.87% B: 5.66%); Agriculture (G: 28.30% B: 39.62%); Building (G: 1.89% B: 9.43%); Tourism (G: 7.55% B: 5.66%)
New firm	Yes (G: 22.64% B: 24.53%); No G: (77.36% B: 75.47%)

Emplo-yee no.	Mean G: 2.32 (σ=3.011);
	Mean B:1.68 (σ=1.47);

Entre-preneur occupa-tion *	Farmers (G: 50.94% B: 41.51%); Retailers (G: 9.43% B: 15.09%); Construction (G: 7.55% B: 11.32%); Elect.engineers, medical (G: 16.98% B: 22.64%); Chemists (G: 15.10% B: 9.43%)
Age	Mean G: 43.28 (σ=10.73); Mean B: 40.55 (σ=8.87)
Business location	Region 1 - G: 37.74% B: 47.17%; Region 2 - G: 28.30% B: 16.87%; Region 3 - G: 11.32% B: 9.43%; Region 4 - G: 22.64% B: 25.53%
Credit request	For the first time G. (88.68% B: 92.45%); Second or third (G: 11.32% B: 7.55%)
Interest pay.*	Monthly (G: 81.13% B: 73.58%); Quarterly (G: 15.09 B: 9.43%); Semi-anually (G: 3.77% B: 16.98%)
Grace period*	Yes (G: 54.72% B: 66.04%); No G: (45.28% B: 33.96%)
Principal paym.*	Monthly (G: 81.13% B: 69.81%); Annualy (G: 18.87% B: 30.19%)
Repay period*	Mean G: 18.78 (σ=6.76) Mean B: 20.19 (σ=6.33)
Interest rate*	Mean G: 13.42 (σ=2); Mean B: 13.02 (σ=1.76)
Credit amount	Mean G: 46370 kunas (σ=31298.84); Mean B: 49403 kunas (σ=33742.11)
Reinv. profit*	G: 50-70%; B: 30-50%
Vision of business *	No (G: 1.89% B: 20.75%); Yes (G: 20.75% B: 3.77); Existing business (G: 77.36% B: 75.47%)
Better then compe-tition in	Quality (G: 35.85% B: 35.85%); Producton (G: 7.55% B: 9.43); Service, price (G: 22.64% B: 5.66%); Reputation (G: 16.98% B: 18.87%); No answer (G: 16.98% B: 30.19%)
Sale level	Local level (G: 60.38% B: 47.17%); Defined customers (G: 9.43 B: 26.42%); One region (G: 11.32% B: 7.55%); Whole country (G: 15.09% B: 15.1%); No ans G: (3.77% B: 3.77%)
Add	No add (G: 16.98% B: 11.32%); All media (G: 22.64% B: 30.19%); Pesonal sale (G: 18.87% B: 1.89%); Internet (G: 3.77% B: 5.67%); No answer (G: 37.74% B: 50.9%)
Aware-ness of competi-tion *	No competition (G: 9.43% B: 18.87%); Broad answer (G: 64.15% B: 52.83%); Defined comp (G: 16.98% B: 5.66%); No answer (G: 9.43 B: 22.64%)

As the output, we use credit scoring in the form of a binary variable with one category representing good applicants and the other one representing bad applicants. An applicant is classified as good if there have never been any payments overdue for 45 days or more, and bad if the payment has at least once been overdue for

46 days or more. The sample consisted of 66% goods and 34% bads. In order to measure the accuracy of models we use hit rates for both NNs and LR. Total hit rate is computed for all correctly classified applicants using the threshold 0.5, as well as individual hit rates for good and bad applicants.

The total sample was divided in sample 1, sample 2a and sample 2b, as presented graphically in Fig. 1.

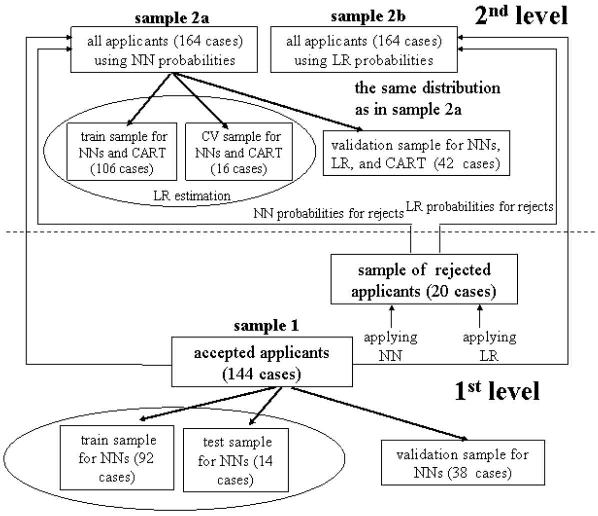
sample 2a
sample 2b
2nd level
all applicants (164 cases) using NN probabilities
all applicants (164 cases) using LR probabilities
the same distribution as in sample 2a
train sample for NNs and CART (106 cases)
CV sample for NNs and CART (16 cases)
validation sample for NNs, LR, and CART (42 cases)
LR estimation
NN probabilities for rejects
LR probabilities for rejects
sample of rejected applicants (20 cases)
sample 1
applying NN
applying LR
accepted applicants (144 cases)
1st level
train sample for NNs (92 cases)
test sample for NNs (14 cases)
validation sample for NNs (38 cases)

Figure 1. Sampling procedure - two levels of experiments

According to the credit scoring methodology suggested by Lewis [11] two credit scoring samples were created. The first sample consisted of the approved applicants only (sample 1). LR and NN models were tested on this sample with the purpose to estimate rejected applicants so that they can be included in the scoring models on the second level of experiments. In such a way the model corresponds to the entire applicants' population and is not biased by previous decision of the bank. The dataset on both levels are divided into the in-sample data (app.75% of data), and the out-of sample data (25% of data) used for final validation. NNs, LR and CART are applied on both samples (2a and 2b) using the same in-sample and out-of-sample data.

4. Results

The first level of experiments extracted the best NN models of each NN algorithm using the previously described nonlinear forward variable selection strategy, as well as the best LR credit scoring model. NN results show that the best total hit rate of 76.3% is obtained by the backpropagation algorithm which also had the best hit rate for bad applicants (84.6%). It is interesting that other three architectures were better at classifying the hit rate for good applicants, but much worse at classifying bad applicants, showing that the key issue for generalization ability of NNs was in recognizing bad applicants. LVQ NN was the worst at performance, unable to classify any bad applicants.

The best extracted first-level LR model regarding standard overall fit measures for logistic regresion (eg. Wald=45.6581 (p=0.0033); Score=77.3657 (p<0.0001)) showed the total hit rate of 83.08%, the goods hit rate of 89.92% and the bads hit rate of 69.69%.

When applied to the first-level validation sample, the best NN classified half of the rejected applicants into good ones, indicating that some of the applicants rejected by the bank should not have been rejected, and therefore corroborating our idea to include the rejected applicants into the whole sample. LR model classified 30% into good and 70% into bad applicants. Since the two methods produced different results on rejected applicants, we were challenged not to rely on NN or LR individually but to include both NN and LR probabilities of rejected applicants into further credit scoring assessments.

When rejected applicants were included in the sample using NN probabilities, the following results were obtained:

Table 2. NN, LR and CART results on the validation sample using NN probabilities for rejected applicants

Model	total hit rate (%)	hit rate of bads (%)	hit rate of goods (%)
Backprop NN, 6-50-2	73.80	53.33	85.19
RBFN, 5-50-2	71.40	73.33	70.37
Probabilistic NN, 10-106-2	83.30	80.00	85.19
LVQ NN, 2-20-2	61.90	13.33	88.89
Logistic regression	57.14	66.67	51.85
CART	66.67	66.67	66.67

As presented in Table 2, the highest NN result, also the overall highest is obtained by probabilistic NN model (total hit rate of 83.3%). This network also showed the highest hit rate in classifying bad applicants (bads hit rate of 80%). Among other NN models, LVQ was the worst at perfomance, unable to classify more than 13.33% of bad applicants, while backpropagation was the second best, followed by RBFN.

Logistic regression results in Table 2 show that the hit rate is higher for bad applicants (66.67%) than for the good ones (51.85%), and that total hit rate is 57.14%.

In contrast to the previous two methodologies, the CART model classifies good and bad applicants with the same accuracy (66.67%), which is higher than the accuracy of LR, but lower than the accuracy of the best NN model.

Table 3. NN, LR and CART results on the validation sample using LR probabilities for rejected applicants

Model	total hit rate (%)	hit rate of bads (%)	hit rate of goods (%)
Backprop NN, 4-50-2	71.40	31.25	92.59
RBFN, 8-50-2	71.40	80.00	66.67
Probabilistic NN, 7-106-2	81.00	93.33	74.07
LVQ NN, 2-20-2	59.50	13.33	85.19
Logistic regression	59.52	68.75	53.85
CART	52.38	100.00	23.08

When NNs use probabilities for rejected applicants obtained by LR, they generally produce lower total hit rates. Table 3 also shows that LR identifies 59.52% of the loans correctly. It can be seen that a higher hit rate is obtained for bad applicants (68.75%) than for the good applicants (53.85%). The CART model gives the worst result among the three methodologies with LR probabilities for the rejected in the data sample (52.38%). However, it is interesting that it classifies correctly 100% of bad applicants, and only 23.08% of the good ones. LR with its own estimation of rejects identifies 59.52% of the applicants correctly and 57.14% with NN estimation for rejects. NNs and CART total hit rates are higher if NN probabilities were used for the rejected applicants, while they both classify bad applicants better if LR probabilities were used.

In order to recognize the best model we computed different measures of association (see eg. [15]) between experimental and estimated values for the applicants in validation sample to be good or bad. The results are given in the Table 4.

Table 4. NN, LR and CART results for measures of association between experimental and estimated values

Measure	NN	CART	LR
phi coefficient	0.66	0.28	0.13
contingency coefficient	0.55	0.27	0.12
Kendal tau-c	0.61	0.27	0.12
Spearman Rank R	0.67	0.28	0.13

According to these measures, the NN model is the best in performance, while the LR is the worst. Considering the fact that the value of phi coefficient for NN model is very high, it can be stated the best that NN model shows a high level of association with experimental data, while the association measures for the LR model are so low that it can be considered unapplicable.

The best overall model (total hit rate 83.30%) is given by the probabilistic NN, with 10 input units and 106 pattern units. For the total hit rate 95% confidence interval is (0,720 - 0,946).

Concerning the variable selection, the best NN model extracted 10 input variables as important. It was proved that both personal and business characteristics are relevant in small business credit scoring systems. Among personal characteristic of entrepreneurs, entrepreneur's occupation was found to be the most important one. Among small business characteristics 4 varibles were found important: clear vision of the business, the planned value of the reinvested profit, main activity of the small business, and awareness of the competition. But, it was also shown that credit program characteristics that are shaped by specific economic conditions of a transitional country are also important in small business credit scoring models. These characteristics are: the way of interest repayment, the grace period, the way of the principal payment, the repayment period, and the interest rate.

5. Conclusion

The paper was aimed to compare the performance of NN, LR and CART decision tree methodologies, as well as to identify important features for the small business credit scoring model on a Croatian dataset. Statistical association measures showed that the best NN model is better associated with data than LR and CART models.. The best NN model significantly outperfomed the LR model, and extracted entrepreneur's personal and business characteristics, as well as credit program characteristics as important features. Since the probabilistic NN algorithm is specifically designed for the problems of classification according to a probability function, recognized and recommended as an efficient classifier in some other areas of application [12], the fact that it performed best in all models indicates that this algorithm could be the one proposed for credit scoring problems of this type.

As guidelines for further research we suggest to extend the dataset with a larger number of cases and a larger number of variables, especially by adding credit bureau data that were not available in our country, but were found relevant for credit scoring by other authors. Secondly, the methodology can be extended with other artificial intelligence techniques such as genetic algorithms, expert systems and others suitable for classification.

6. References

[1] Altman EI, Marco G, Varetto F. Corporate Distress Diagnosis: Comparison Using Linear Discriminant Analysis and Neural Networks (the Italian Experience). Journal of Banking and Finance 1994; 18: 505-529.

[2] Apte C, Weiss S. Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems 1997; 13 (2): 197-210.

[3] Arriaza, B.A. Doing Business with Small Business. Business Credit Nov/Dec 1999; 101(10): 33-36.

[4] Berry MJA., Linoff G. Data Mining Techniques for Marketing, Sales and Customer Support. New York: John Wiley & Sons; 1997.

[5] Fahrmeier L, Tutz G. Multivariate Statistical Modeling Based on Generalized Linear Models. Berlin: Springer, 2001.

[6] Feldman R. Small Business Loans, Small Banks and a Big Change in Technology Called Credit Scoring. Region 1997; 11(3): 18-24.

[7] Frame WS, Srinivasan A, Woosley L. The Effect of Credit Scoring on Small Business Lending. Journal of Money, Credit and Banking 2001; 33(3): 813-825.

[8] Friedland M. Credit Scoring Digs Deeper Into Data. Credit World, 1996: 84(5): 19-24.

[9] Galindo J, Tamayo P. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics 2000; 15: 107-143.

[10] Hand DJ, Henley WE. Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of Royal Statistical Society A 1997: 160: 523-541.

[11] Lewis EM. An Introduction to Credit Scoring, San Rafael: Fair Isaac and Co. Inc; 1992.

[12] Masters T. Advanced Algorithms for Neural Networks, A C++ Sourcebook. New York: John Wiley & Sons; 1995.

[13] NeuralWare. Neural Computing, A Technology Handbook for NeuralWorks Professional II/Plus and NeuralWorks Explorer, NeuralWare. Pittsburgh: Aspen Technology, Inc; 1998.

[14] Patterson DW. Artificial Neural Networks, New York: Prentice Hall, 1995.

[15] Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures. Washington D.C.: CRC Press; 1997

[16] Refenes AN, Zapranis A, Francis G. Stock Performance Modeling Using Neural Networks: A Comparative Study with Regression Models. Neural Networks 1997; 7(2): 375-388.

[17] Skare M. Ogranicenja razvoja malih i srednjih poduzeca u zemljama tranzicije. Gospodarska politika, srpanj, 2000.

[18] West D. Neural Network Credit Scoring Models. Computers & Operations Research. 2000; 27: 1131-1152.

[19] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation. San Francisco: Morgan Kaufman Publishers; 2000.

[20] Yobas MB, Crook JN, Ross P. Credit Scoring Using Evolutionary Techniques. IMA Journal of Mathematics Applied in Business & Industry 2000; 11: 111-125.